Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 195 (2015) 1865 – 1871

Procedia
Social and Behavioral Sciences

World Conference on Technology, Innovation and Entrepreneurship

# The Effect of Variations in Micro-Components of Domestic Water Consumption Data on The Classification Of Excessive Water Usage

Nor Salyana Mohd Salleh[a], Khairul A. Rasmani[b] and Nur Izzah Jamil[b*]

[a]*Faculty of Science, Universiti Teknologi Malaysia, Johor, Malaysia*
[b]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Seremban Campus, N. Sembilan, Malaysia*

**Abstract**

Several studies have been conducted to explore the potential of using consumer data for the estimation of domestic water consumption. However, it is not an easy task to collect and record consumer data that precisely represent daily, weekly or monthly household water usage. This paper investigates the effect of variations in water consumption data on the classification of domestic water usage levels. Two datasets were used in this study. The first dataset consists of ten predictive variables related to household water usage. The second dataset was generated based on the first dataset where four generic features were created to represent water consumption based on four categories of activities related to water usage. Selected classification algorithms were used for classification task. The findings show that variations in consumer data have very little effect on classification outcomes suggesting that data collected from consumer suitable to be used for predicting excessive domestic water usage.


Peer-review under responsibility of Istanbul Univeristy.

*Keywords:* Consumer data; domestic water consumption; classification

## 1. Introduction

Clean water is used in a household for various important reasons. Typically clean water is used for personal use related to hygiene, health, cooking and food preparation, washing clothes and other related indoor and outdoor activities such as washing car and gardening (Arbués, García-Valiñas, & Martínez-Espiñeira, 2003; Corbella &

* Corresponding author. tel.: +606-6342000; fax: +606-6335813.
*E-mail address:* khairulanwar@ns.uitm.edu.my


Peer-review under responsibility of Istanbul Univeristy.
doi:10.1016/j.sbspro.2015.06.422

Pujol, 2009; Gleick, 1996; Schleich & Hillenbrand, 2009). Collecting consumer water consumption data can be very challenging as the levels of water usage is expected to be varied between individuals. Although attempts to use consumer water consumption data to classify levels of water usage have been investigated (Ismail et al., 2012; Mohd Rashid et al., 2014), it can be observed that lack of study have been conducted on the effect of variations in consumer water consumption data on the prediction outcomes produced by classifications methods.

The main objective of the research presented in this paper is to investigate the effects of variations in the data collected from consumers containing micro-components of household water usage on classification of excessive water usage performs using classification algorithms. Variations in water consumption data collected from the consumer is expected to be very high as data collected from the consumer is not normally based on exact measurement but purely based on estimation by the consumer. This is particularly true as collecting water consumption data from the consumer can be very tricky as the number of times people conducting activities related to water consumption are not easily recorded whether on daily, weekly or monthly basis. Collecting data using 'diary recording' (O'Toole, Sinclair, & Leder, 2009) may be implemented for the purpose of conducting a research study but obviously is not practical to be practiced by consumers. Additionally, in estimating total water consumption based on micro-components of water consumption, the amount of water per activities need to be chosen very carefully as many studies found that the estimated water usage differs from one place to another (Aquaterra, 2008).

Domestic per capita water consumption is the amount of water consumed per person for the purposes of ingestion, hygiene, cooking, washing of utensils and other household purposes including garden uses. Hence, the total water usage not only affected by individual socio-economic background (Schleich & Hillenbrand, 2009) but also by some other factors such as the religious and cultural belief (Smith & Ali, 2006). Some other insignificant factors may as well contribute such as the age of people living in the household as study suggest that young and old people use the water less than the average (Schleich & Hillenbrand, 2009). Urban and non-urban factor has also been identified to contribute to the difference in domestic water consumption. A more complicated factor is the inter-relationship of individual domestic water consumption with weather, geographical location or even the size of the properties (Balling, Gober, & Jones, 2008; Schleich & Hillenbrand, 2009).

Certain technical aspects such as water flow rate (Corona-Nakamura, Ruelas, Ojeda-Magana, & Andina, 2008) and types of water appliances will also contribute to the correctness of the estimation of domestic water consumption by the consumers. Different amount of water usage for a specific activity is expected in different place and time setting due variation of water flow rate in a household. The level of water usage also depends on appliances used for the water related activities. For example, in the past decade a number of technical measures on appliance have been developed to reduce domestic water consumption (Terpstra, 1999). Obviously some of these aspects may be unknown by the user.

From the literature, it can be observed that the main source of variations in micro-components of domestic water consumption data comes from various aspects. The most significant aspects are the data collection and classification methods, followed by socio-economic, socio-demographic and environmental factors while the technical aspects related to water supply and consumption by water appliances may as well contribute to the variation in the estimation of water consumption by the consumers. It can be concluded that variations in micro-components of domestic water consumption data is unavoidable.

## 2. Experimental Set-up

The main objective of the research presented in this paper is to investigate the effect of data variations on classification outcomes. In order to achieve this objective, the dataset originally used in a previous research (Mohd Hanif, Rasmani, & Mohamed Ramli, 2013) has been used with modification. For the purpose of conducting the experiments, two sets of data were prepared. The first set of data (labelled as DF0) consists of ten predictive

variables whereas the second set of data (labelled as DS0) consists of four generic features created purposely to represent water usage based on four categories of activities related to water usage. The generated features are individual water usage, water use for food preparation and cooking related activities, water use for washing clothes and water for other activities such as car washing and gardening. For the purpose of conducting this research the estimated amount of water used per activity has been fixed and the values are not referring on any specific research outcome. Note that certain variables in the datasets recorded zero value such as car washing and washing clothes by hand which indicate that these activities are uncommon. The classification outcomes of each instance in the dataset were based on the research outcomes reported in (Mohd Hanif et al., 2013) where classical Hotelling $T^2$ control chart were used to identify potential excessive water usage cases in which the classification outcomes were categorized as "Likely" and "Unlikely".

For the purposes of conducting the experiments, different datasets need to be created to represent variations in consumer water consumption data. Based on the two original datasets mentioned above, additional seven sets of simulated data were created. The first data set was created by adding each of the predictive variable by a random generated value up to 10% from the original value. Other six additional datasets were created using the same method, but the value were increased further 10% each time. Note that the addition of the original value with additional 10% is to represent the situation that some households used water 10% higher than usual. Hence, besides the original datasets, seven simulated datasets were prepared to represent each category. For the first category of datasets (data with ten original predictive variables), the eight sets of consumer water consumption data were labelled as DF0, DF1, DF2, …, DF7 while the second category of datasets (data with four generic predictive variables) were labelled as DS0, DS1, DS2, …, DS7, respectively.

To compare the classification outcomes between datasets, selected fuzzy and non-fuzzy classification algorithms available in WEKA Machine Learning Software (Witten, Frank, & Hall, 2011) were used. This software contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Therefore, besides using several classification algorithms already available in WEKA Version 3.6, several new algorithms were also implemented in WEKA platform. The main assumption in using the algorithms is that the consumer water consumption datasets are suitable to be used for the selected classification algorithms. The classification accuracy calculated in the analysis are based on 10-fold cross validation of each algorithm on each dataset. The effect of variation on classification outcomes is measured based on comparison of the classification accuracies. Note that the aim of this study is not to compare the performance between fuzzy with non-fuzzy approaches but the results were presented separately to observe if there are any significant differences between these two approaches. Besides that, the fuzzy membership functions used for fuzzy methods were not in any way optimized to get better classification outcomes.

## 3. Experimental Results and Discussions

To achieve the objective outlined in this study, datasets with different variations need to be created. Analysis to compare the variation between datasets were conducted using Tukey's method available in MINITAB software for the datasets with ten predictive attributes and four generic attributes, respectively. Table 1 and Table 2 show the experimental results of the multiple comparisons.

Table 1 represents comparison based on the MINITAB output for the household water usage data based on Tukey 95% simultaneous confidence intervals which conclude that the original data (DF0) is statistically different from data DF3, DF4, DF5, DF6 and DF7 for variables labelled as 'A1', 'A2', 'A3' and 'A5'. Results obtained from the analysis also show that the original data (DF0) is statistically different from data DF4, DF5, DF6 and DF7 for variables labelled as 'A4', 'A6', 'A8' and 'A9'. The analysis also indicates that the original data is statistically different from data DF5, DF6 and DF7 for variable 'A8' and statistically different from data DF6 and DF7 for variable 'A10'.

Table 2 represents the summary of 95% Tukey simultaneous confidence interval from the MINITAB output for the household water usage data that contains four generic predictive attributes. Based on Tukey 95% simultaneous confidence intervals it can be concluded that the original data (DS0) is significantly different from data DS3, DS4, DS5, DS6 and DS7 for variable 'G1 - Individual usage'. Results also show that the original data is significantly different from data DS4, DS5, DS6 and DS7 for variables 'G2 – Washing clothes' and 'G3 – Food preparation'. Analysis also indicates that the original data is significantly different from data DS5, DS6 and DS7 for variable 'G4 – Outdoor and other activities'.

Table 3 and Table 4 present the classification accuracy obtained from each classification method for each dataset. For fuzzy classification methods, the results clearly show that the average classification accuracy between the datasets is very close where the highest classification accuracy (92.10%) obtained for the original data (DF0) while the lowest average classification accuracy (91.75%) obtained for data labelled as DF2. Most algorithms also performs consistently throughout the eight datasets. The similar results can be observed for non-fuzzy classification algorithms.

Table 5 and Table 6 present comparison of classification accuracy obtained from the second category of datasets which consist of four generic predictive variables. The objective in conducting this experiment is to provide additional evidence to support or reject findings obtained using the first category of datasets. The experimental results show that the highest average classification accuracy (92.31%) produced by fuzzy classification methods obtained from the original data (DS0) while the lowest average classification accuracy (90.23%) obtained by data labelled as DS7. Very similar results can be observed from the classification outcomes obtained using non-fuzzy classification algorithms.

Table 1. Comparison of the original data (DF0) with the simulated data (Ten variables)

| Variable | DF1 | DF2 | DF3 | DF4 | DF5 | DF6 | DF7 |
|---|---|---|---|---|---|---|---|
| A1 | NS | NS | S | S | S | S | S |
| A2 | NS | NS | S | S | S | S | S |
| A3 | NS | NS | S | S | S | S | S |
| A4 | NS | NS | NS | S | S | S | S |
| A5 | NS | NS | S | S | S | S | S |
| A6 | NS | NS | NS | S | S | S | S |
| A7 | NS | NS | NS | NS | S | S | S |
| A8 | NS | NS | NS | S | S | S | S |
| A9 | NS | NS | NS | S | S | S | S |
| A10 | NS | NS | NS | NS | NS | S | S |

*S- Significant; NS- Not Significant*

Table 2. Comparison of the original data (DS0) with the simulated data (Four variables)

| Variable | DS1 | DS2 | DS3 | DS4 | DS5 | DS6 | DS7 |
|---|---|---|---|---|---|---|---|
| G1 | NS | NS | S | S | S | S | S |
| G2 | NS | NS | NS | S | S | S | S |
| G3 | NS | NS | NS | S | S | S | S |
| G4 | NS | NS | NS | NS | S | S | S |

*S- Significant; NS- Not Significant*

Table 3. Comparison of Classification Accuracy Obtained using Selected Fuzzy Classification Methods (Ten Variables)

| Fuzzy Methods | Classification accuracy (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | DF0 | DF1 | DF2 | DF3 | DF4 | DF5 | DF6 | DF7 |
| Fuzzy Rough NN1 | 92.94 | 92.03 | 92.03 | 92.26 | 93.39 | 92.71 | 91.34 | 91.34 |
| Fuzzy Rough NN2 | 91.80 | 91.80 | 91.57 | 91.80 | 91.80 | 91.80 | 92.03 | 91.34 |
| Fuzzy K-NN | 92.48 | 92.71 | 92.71 | 93.17 | 91.80 | 92.03 | 93.17 | 92.71 |
| Fuzzy O K-NN | 92.03 | 92.26 | 92.03 | 92.26 | 93.17 | 92.26 | 92.26 | 93.62 |
| DNNC | 93.17 | 92.71 | 92.71 | 92.94 | 92.94 | 92.71 | 93.17 | 92.71 |
| FRRI | 93.62 | 94.31 | 92.71 | 94.31 | 94.53 | 94.08 | 93.39 | 92.94 |
| VQ FRRI | 93.17 | 93.85 | 92.71 | 93.39 | 93.39 | 93.17 | 93.62 | 93.17 |
| VQ FNN | 91.80 | 91.80 | 91.57 | 91.80 | 91.80 | 91.80 | 92.03 | 91.34 |
| Average (%) | 92.63 | 92.68 | 92.26 | 92.74 | 92.85 | 92.57 | 92.63 | 92.40 |

Table 4. Comparison of Classification Accuracy Obtained using Selected Non-Fuzzy Classification Methods (Ten Variables)

| Non-Fuzzy Methods | Classification accuracy (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | DF0 | DF1 | DF2 | DF3 | DF4 | DF5 | DF6 | DF7 |
| RBF Network | 96.81 | 96.58 | 96.36 | 96.36 | 94.53 | 96.13 | 94.31 | 95.90 |
| Multilayer Perceptron | 94.31 | 94.08 | 94.31 | 94.76 | 93.85 | 95.22 | 94.08 | 95.67 |
| Simple Logistic | 94.99 | 94.99 | 94.53 | 95.44 | 94.53 | 95.90 | 95.22 | 95.44 |
| LMT | 94.99 | 94.99 | 95.53 | 95.44 | 94.53 | 95.90 | 95.22 | 95.44 |
| Rotation Forest | 94.76 | 95.44 | 95.90 | 94.31 | 94.31 | 94.76 | 94.08 | 95.22 |
| Logistic | 94.76 | 95.22 | 95.90 | 95.22 | 94.99 | 94.99 | 94.99 | 94.99 |
| Multi-Class Classifier | 94.76 | 95.22 | 95.90 | 95.22 | 94.99 | 94.99 | 94.99 | 94.99 |
| Naïve Bayes Simple | 93.17 | 92.48 | 92.48 | 91.80 | 92.03 | 92.48 | 91.34 | 92.94 |
| SMO | 95.44 | 95.67 | 95.67 | 95.22 | 94.76 | 94.99 | 95.44 | 94.31 |
| Average (%) | 94.89 | 94.96 | 95.18 | 94.86 | 94.28 | 95.04 | 94.41 | 94.99 |

From the results obtained in both experiments using the first and second category of datasets, it can be concluded that the classification outcomes are very close. These results are consistent in all four cases as presented in Table 3 – Table 6. For the first experiment, the comparison of the classification outcomes between the original data (DF0) with the data labelled as DF6 and DF7, the different in classification accuracy is very small although the multiple comparison between all variables show the variation between variables are statistically significant. The same observation can be seen for the second experiment where comparison (presented in Table 5 and Table 6) on the difference in average classification outcomes between the original data (DS0) and data labelled as DS5, DS6 and DS7 are very small although the multiple comparison between all variables show the variation between variables are statistically significant.

Table 5. Comparison of Classification Accuracy Obtained using Selected Fuzzy Classification Methods (Four Variables)

| Fuzzy Methods | Classification accuracy (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | DS0 | DS1 | DS2 | DS3 | DS4 | DS5 | DS6 | DS7 |
| Fuzzy Rough NN1 | 93.85 | 93.85 | 93.17 | 93.85 | 92.03 | 92.71 | 91.80 | 90.43 |
| Fuzzy Rough NN2 | 93.17 | 92.48 | 92.94 | 93.17 | 92.48 | 92.26 | 92.03 | 91.57 |
| Fuzzy K-NN | 93.85 | 94.31 | 94.53 | 94.31 | 93.62 | 93.85 | 92.03 | 92.03 |
| Fuzzy O K-NN | 92.71 | 91.57 | 92.26 | 91.80 | 91.80 | 91.80 | 91.34 | 90.43 |
| DNNC | 93.39 | 93.39 | 93.17 | 92.94 | 92.48 | 91.57 | 92.03 | 91.57 |
| FRRI | 91.80 | 92.03 | 91.34 | 91.80 | 93.17 | 91.80 | 88.15 | 89.52 |
| VQ FRRI | 91.12 | 91.12 | 91.34 | 91.80 | 91.80 | 90.89 | 90.66 | 90.89 |
| VQ FNN | 93.17 | 92.48 | 92.94 | 93.17 | 92.48 | 92.26 | 92.03 | 91.57 |
| Average (%) | 92.88 | 92.65 | 92.71 | 92.86 | 92.48 | 92.14 | 91.26 | 91.00 |

Table 6. Comparison of Classification Accuracy Obtained using Selected Non-Fuzzy Classification Methods (Four Variables)

| Non-Fuzzy Methods | Classification accuracy (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | DS0 | DS1 | DS2 | DS3 | DS4 | DS5 | DS6 | DS7 |
| RBF Network | 94.08 | 93.85 | 93.62 | 93.17 | 93.39 | 92.94 | 93.62 | 92.48 |
| Multilayer Perceptron | 94.53 | 94.31 | 94.08 | 93.17 | 94.08 | 93.39 | 93.85 | 92.03 |
| Simple Logistic | 94.99 | 94.53 | 94.76 | 94.53 | 94.53 | 93.85 | 94.53 | 93.62 |
| LMT | 94.99 | 94.53 | 94.31 | 94.53 | 94.31 | 93.85 | 95.53 | 93.62 |
| Rotation Forest | 94.08 | 94.31 | 94.76 | 92.48 | 93.62 | 93.85 | 94.31 | 93.39 |
| Logistic | 94.99 | 94.76 | 94.31 | 94.31 | 93.39 | 94.31 | 94.31 | 93.62 |
| Multi-Class Classifier | 94.99 | 94.76 | 94.31 | 94.31 | 93.39 | 94.31 | 94.31 | 93.62 |
| Naïve Bayes Simple | 92.94 | 94.08 | 93.39 | 93.17 | 93.17 | 93.62 | 93.17 | 91.80 |
| SMO | 92.94 | 92.94 | 92.03 | 92.94 | 92.94 | 90.89 | 91.12 | 91.12 |
| Average (%) | 94.28 | 94.23 | 93.95 | 93.62 | 93.65 | 93.45 | 93.86 | 92.81 |

## 4. Conclusions

This paper has presented a study on the effect of variations in micro-components of domestic water consumption data on the classification of excessive residential water usage. The findings showed that there were little evidence to suggest that variation in water consumption data have significant impact on the classification outcomes. This is true not only based on the result obtained using fuzzy classification approaches but also based on the results obtained using non-fuzzy classification approaches. Although these results seem very promising, more studies should be carried out to confirm the finding reported in this paper.

## Acknowledgements

This research work is funded by the Ministry of Education, Malaysia under the Fundamental Research Grant Scheme (FRGS) with reference number 600-RMI/ TD/FRGS 5/3 (1/2015). The authors also would like to thank the Research Management Institute (RMI), Universiti Teknologi MARA, Malaysia.

## References

Aquaterra. (2008). International comparisons of domestic per capita consumption. Bristol: Environment Agency.

Arbués, F., García-Valiñas, M. a. Á., & Martínez-Espiñeira, R. (2003). Estimation of residential water demand: a state-of-the-art review. *The Journal of Socio-Economics, 32*(1), 81-102.

Balling, R. C., Gober, P., & Jones, N. (2008). Sensitivity of residential water consumption to variations in climate: An intraurban analysis of Phoenix, Arizona. *Water Resources Research, 44*.

Corbella, H., & Pujol, D. i. (2009). What lies behind domestic water use?: A review essay on the drivers of domestic water consumption. *Boletín de la Asociación de Geógrafos Españoles, 50*, 297-314.

Corona-Nakamura, M. A., Ruelas, R., Ojeda-Magana, B., & Andina, D. (2008). Classification of domestic water consumption using an ANFIS model. *World Automation Congress*, 1-9.

Gleick, P. H. (1996). Basic Water Requirements for Human Activities: Meeting Basic Needs. *Water International, 21*(2), 83-92.

Ismail, N. F., Rasmani, K. A., Shahari, N., Rashid, N. R. M., Hanif, H. M., & Noh, N. A. M. (2012, 14-15 Aug. 2012). *Prediction of residential households' water leakage using consensus method.* Paper presented at the 2nd International Conference on Uncertainty Reasoning and Knowledge Engineering (URKE).

Mohd Hanif, H., Rasmani, K. A., & Mohamed Ramli, N. (2013). *Detecting residential household water leakage and wastage using the Individual Hotelling $T^2$ control chart.* Paper presented at the IJAS American-Canadian International Conference on Academic Disciplines.

Mohd Rashid, N. R., Rasmani, K. A., Shahari, N., Ismail, N. F., Mohd Hanif, H., & Noh Md, N. A. (2014). Prediction of Domestic Water Leakage Based on Consumer Water Consumption Data. *Advanced Science Letters, 20*(1), 344-347.

O'Toole, J., Sinclair, M., & Leder, K. (2009). Collecting household water usage data: telephone questionnaire or diary? *BMC Medical Research Methodology, 9*(1), 72.

Schleich, J., & Hillenbrand, T. (2009). Determinants of residential water demand in Germany. *Ecological Economics, 68*(6), 1756-1769.

Smith, A., & Ali, M. (2006). Understanding the impact of cultural and religious water use. *Water and Environment Journal, 20*, 203-209.

Terpstra, P. M. J. (1999). Sustainable water usage systems: Models for the sustainable utilization of domestic water in urban areas. *Water Science and Technology, 39*(5), 65-72.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.): Morgan Kaufmann.